Comparing Manual and Automatic UMRs for Czech and Latin

Jan Štěpánek and Daniel Zeman and Markéta Lopatková and
Federica Gamba and Hana Hledíková
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské nám. 2/25, 118 00 Prague 1, Czechia
{stepanek,zeman,lopatkova,gamba,hana.hledikova}@ufal.mff.cuni.cz

Abstract

Uniform Meaning Representation (UMR) is a semantic framework designed to represent the meaning of texts in a structured and interpretable manner. In this paper, we evaluate the results of the automatic conversion of existing resources to UMR, focusing on Czech (PDT-C treebank) and Latin (LDT treebank). We present both quantitative and qualitative evaluations based on a comparison between manually and automatically generated UMR structures for a sample of Czech and Latin sentences. The findings indicate comparable results of the automatic conversion for both languages. The key challenges prove to be the higher level of semantic abstraction required by UMR and the fact that UMR allows for capturing semantic structure in multiple ways, potentially with varying levels of granularity.

1 Introduction

The challenge of representing meaning has been fascinating linguists, philosophers, and cognitive scientists for centuries. Traditional semantic frameworks—such as truth-conditional semantics (e.g., Davidson, 1967), frame semantics (e.g., Baker et al., 1998; Fillmore et al., 2002), and cognitive semantics (e.g., Langacker, 1987; Croft and Cruse, 2004)—aimed to formalize how meaning is constructed, interpreted, and communicated.

Recent advances in natural language processing have been driven by large language models. These models excel at downstream tasks such as text generation and translation. However, given their unclear interpretability—as they rely on statistical patterns rather than true semantic or logical understanding—they do not answer the essential questions about meaning representation.

Thus, symbolic approaches remain central to efforts to search for precise, inference-capable meaning representations. *Uniform Meaning Representation* (UMR), the fundamentals described by van Gysel et al. (2021), is one of the responses to this interest. We build on this initiative and test the approach for representing Czech and Latin—inflected languages with rich morphology and free word order representing information-structural features (such as topic-focus articulation and discourse dynamics) rather than syntactic relations. The results of our effort could thus provide valuable insight for the UMR community.

Creating data from scratch is extremely time-consuming and requires highly trained annotators with extensive expertise. That's why we aim to take advantage of the richly annotated datasets already available for the two languages, and investigate the possibility of their (semi-)automatic conversion to the UMR framework. Namely, we rely on the PDT-C corpus[1] (Hajič et al., 2024a) for Czech and on a subset of the Latin Dependency Treebank (LDT)[2] (Bamman and Crane, 2006) for Latin. Both are annotated using the same PDT annotation scenario, thus supporting the same conversion process. A similar approach has proved to be advantageous for English—as described by Bonn et al. (2023b), who created the extensive English UMR corpus (Bonn et al., 2025) from structures used in *Abstract Meaning Representation*, the UMR predecessor. Full conversion is not always feasible, but even partial results are highly beneficial, as shown by Buchholz et al. (2024) and Gamba et al. (2025).

The paper presents a comparison of (a small sample of) double-annotated UMR data, that is, the data with manually created UMR structures and their counterparts automatically converted from the PDT-C and LDT corpora, respectively. First, we briefly describe the UMR and PDT-C approaches and the available automatic conversion (§ 1.1, 1.2, and 1.3, respectively) and the Czech

[1] http://hdl.handle.net/11234/1-5813

[2] https://itreebank.marginalia.it/

Proceedings of the 6th International Workshop on Designing Meaning Representations, pages 1–12
August 4, 2025. ©2025 Association for Computational Linguistics

and Latin UMR data (§ 2). § 3 introduces the way we compare the structures and brings a quantitative comparison. A qualitative analysis follows in § 4. § 5 then summarizes the results and discusses further work.

1.1 Uniform Meaning Representation

Uniform Meaning Representation (see esp. van Gysel et al., 2021; Bonn et al., 2023b, 2024) is a semantic framework designed to represent the meaning of texts in an interpretable way, elaborating the (originally English-centered) *Abstract Meaning Representation* (Banarescu et al., 2013; Wein and Bonn, 2023). UMR's graph-based sentence-level representation abstracts from the overt sentence syntax; in particular, it encodes the frame-based predicate-argument structure of all eventive concepts, including their aspectual information. In addition, UMR models semantic relations that cross sentence boundaries, such as coreference, temporal chains, and epistemic modality, which makes it possible to interpret context and discourse more effectively.[3] Its applicability has been demonstrated on a sample of data from English, Chinese, and four low-resource American languages (Bonn et al., 2023a).

1.2 PDT: Deep syntactic representation

Both treebanks that we use as our source data, Czech PDT-C and Latin LDT, provide representation at the so-called deep syntactic layer (also tectogrammatical or t-layer; see esp. Sgall et al., 1986; Hajič et al., 2020 for Czech and Passarotti, 2014; Gonzalez Saavedra and Passarotti, 2014 for Latin). The core of this dependency-oriented representation is formed by the predicate-argument structure (valency) and other deep syntactic relations. This core structure is enriched with meaning-relevant morphological information (number and gender for nouns; tense, aspect, modality for verbs), topic-focus articulation, and coreference annotation.[4]

In contrast to UMR, the PDT scenario concentrates on linguistically structured meaning; as such, it more or less directly refers to the annotated text. Thus, this scenario is less abstract than UMR—which presents the main obstacles to the automatic conversion (as will be discussed below).

A more thorough comparison of the two approaches, envisaging the possibility of the automatic PDT-C to UMR conversion, can be found in Lopatková et al. (2024).

1.3 PDT to UMR automatic conversion

Here we work with the first attempt to automatically convert PDT structures to UMR structures, as described in Lopatková et al. (2025). Let us stress that this conversion is partial—it covers only selected phenomena pertaining to the sentence-level annotation (esp. structure of the graph, labeling of nodes and relations, PropBank-like argument structure for verbs, and selected attributes); in addition, intra-sentential coreference relations are identified.

The conversion procedure recursively traverses the PDT-C tree (namely the t-structure), and incrementally builds the corresponding UMR graph. Each node and edge are analyzed to determine necessary structural and labeling changes, as well as the addition of UMR attributes.

- In this stage, *structural transformations* are a key part of the process. These typically arise from handling coreference (merging pronouns with their referents, reentrancies, inverse roles), coordination (esp. representing conjuncts and their shared dependents in a UMR-adequate structure), relative clauses (merging referential nodes and linking them semantically), and control or raising verbs (merging arguments across predicates), as sketched by Lopatková et al. (2024, 2025).
- Changes in *nodes labeling* reflect the shift from deep syntactic elements of PDT-C (identified as t-lemmata) to UMR concepts (entities, states, and processes).
- For *edges labeling*, deep syntactic roles of PDT-C are converted to UMR semantic relations, using (i) verb-specific mapping of arguments (whenever available, Hajič et al., 2024b) and (ii) default mapping of arguments and adjuncts (Lopatková et al., 2025).
- UMR nodes are enriched with selected UMR attributes, namely aspect, degree, polarity, quant, refer-number, and refer-person (Lopatková et al., 2025).
- *Nodes alignment* is gained from PDT-C.

In the following, we concentrate on evaluating the quality of conversion for the aforementioned phenomena. We exclude UMR attributes not listed above (i.e., wiki, modal-strength, mode, polite, and

[3] The UMR 0.9 specification as available here: https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md

[4] For the full PDT-C documentation, see https://ufal.mff.cuni.cz/pdt-c/documentation.

	corpus	sentences	tokens	PDT / LDT nodes	UMR nodes (manual)	UMR nodes (automatic)
Czech	PDT	25	467	378	375	349
	PDTSC	50	374	321	442	305
	PCEDT	16	474	400	307	327
	total	91	1315	1099	1124	981
Latin	LDT	50	889	928	773	865

Table 1: Statistics for both manually and automatically annotated data.

quote), as well as all phenomena represented in the document-level annotation.[5]

2 Double-annotated UMR Data

2.1 Czech UMR data

The PDT-C corpus offers a large volume of Czech data spanning various genres. We selected a sample of six files from its development data for manual annotation. This sample covers key genres presented in PDT-C (written texts in both general journalistic and technical styles, as well as spoken data). Another selection criterion was that the files include specific linguistic phenomena where we anticipate problems during the conversion (e.g., not overtly expressed entities or events, selected types of special constructions, coordinated structures, complex coreferential chains, negation). Specifically, the selected texts are as follows:

- 25 sentences (2 documents) from the core PDT[6] subcorpus (Czech newspaper texts from 1992-94);
- 50 sentences from the PDTSC[7] subcorpus (spontaneous dialogs);
- 16 sentences (out of 37 sentences, 2 documents) from the Czech part of the PCEDT[8] subcorpus (Czech translations of the Penn Treebank-WSJ texts).

Table 1 provides more detailed statistics. It reveals that the WSJ texts from PCEDT are more complex (especially compared to spontaneous dialogs from PDTSC); thus, despite the lower number of PCEDT sentences, the sample data selected for manual annotation provide relatively balanced coverage of the genres represented in the corpus.

A small portion of the data (21 sentences with 255 tokens from PDT and PDTSC) were annotated by two human annotators in parallel; these data were used to estimate inter-annotator agreement (Table 2).

2.2 Latin UMR data

The corpus utilized in this study corresponds to a portion of the LDT as provided by the *Index Thomisticus Treebank* project[9] (Passarotti, 2019). Compared to the original version, this subset was refined at the syntactic layer and annotated from scratch at the semantic-pragmatic layer. It includes the entire *De coniuratione Catilinae* ‘Conspiracy of Catiline’ by Sallust along with excerpts from the works of Caesar and Cicero. For this work, we focus specifically on Sallust and select the first 50 sentences of his work, corresponding to the first five (out of 61) chapters of the text. We select these sentences as they are already part of the UMR 2.0 release. Table 1 provides basic data statistics.

3 Comparison: Global Perspective

3.1 Metrics for graph comparison

Quantitative comparison of semantic graphs is a non-trivial task because two representations of the same sentence may differ in the number of nodes, and the node identifiers (variables) typically differ, too. It is thus not obvious which nodes should be taken as corresponding to each other. If we can find the optimal node mapping between the two graphs, the rest of the task is easy. Properties of the graph can be expressed as a set of triples (x, y, z), where x is a node (now identifiable in both graphs), y is a name of a relation or an attribute, and z is another node (child node of the relation) or the value of the attribute. Similarity of two graphs can be expressed as the F_1 score of the triples.

UMR is a successor to AMR, and for AMR, the *smatch* metric (Cai and Knight, 2013) has emerged as the de-facto standard. It defines as optimal the mapping that maximizes F_1 of the resulting triples;

[5] The Czech and Latin UMR data described and compared in the paper are available through the Lindat repository, see http://hdl.handle.net/11234/1-5951.

[6] https://ufal.mff.cuni.cz/pdt3.5

[7] https://ufal.mff.cuni.cz/pdtsc2.0

[8] https://ufal.mff.cuni.cz/pcedt2.0/

[9] https://itreebank.marginalia.it/

UMR node mapping:					
Anot1 nodes	Anot2 nodes	mapped	recall	precision	F$_1$
228	221	215	94%	97%	96%

Concept and relation comparison (only mapped nodes):*					
Anot1 triples	Anot2 triples	match	recall	precision	F$_1$
633	644	595	94%	92%	93%

Concept and relation comparison:**					
Anot1 triples	Anot2 triples	match	recall	precision	*ju:mæʧ* = F$_1$
663	659	595	90%	90%	90%

Table 2: Manually double-annotated UMRs: quantitative comparison for Czech (PDT+PDTSC). (* Unmapped nodes are ignored. ** Unmapped nodes all counted as incorrect.)

the *smatch* algorithm employs hill-climbing with restarts to find an approximate solution to the optimization problem.

An alternative node mapping algorithm, called *AnCast*, has been proposed specifically for UMR (Sun and Xue, 2024). It has been shown to be more efficient and more accurate than *smatch*. The authors also define a number of partial metrics, such as Concept F_1 and Labeled Relation F_1, which improve interpretability of the results.

One of the improvements of UMR over AMR is that UMR annotation includes alignment of nodes to surface tokens. *Smatch* does not have the notion of word alignment; *AnCast* can use it if available, but it can work without it, too. Nevertheless, *AnCast's* ability to exploit alignment is limited. The token–node alignments can be $M : N$, with a node potentially mapped to a discontinuous set of tokens, while *AnCast* can currently process only continuous alignments. *AnCast* also compares concepts of the nodes to be mapped, and it tries to assess concept similarity rather than identity, although in a restricted manner. To achieve similarity > 0, one concept lemma must be substring of the other. This would recognize similarity between e.g., *fry* and *stir-fry*, but not between Czech *volit* 'to vote' and nominalized *volba* 'election'.

Both *smatch* and *AnCast* will map as many nodes as possible. If one of the graphs has more nodes that the other, remaining nodes will stay unmapped. If the graphs have the same number of nodes, every node will be mapped to a node in the other graph, even if they are clearly unrelated. This may occasionally improve the score when a random attribute occurs in both nodes, but it blurs the interpretation of the score. More importantly, we also want to use the mapping to eye-ball disagreement between annotators, and maximal node mapping is not helpful for that purpose. Therefore, we employ a third mapping algorithm called *ju:mæʧ*, which primarily maps nodes aligned to the same word(s), and for nodes without word alignment (which are a minority in UMR graphs) requires concept identity. As with *smatch* and *AnCast*, we assess similarity of other node attributes if needed to get a symmetric one-to-one mapping. An example comparing *ju:mæʧ* and *smatch* mappings is given in Appendix A.

Note that all scores in the present paper evaluate only the sentence-level graphs in UMR. The document-level relations (modal and temporal annotation, coreference) could be evaluated as triples using the same node mapping, but the current evaluation scripts do not support it.

3.2 Quantitative comparison

Comparison of manually double-annotated Czech data. First, to gain insight into the problem, we quantitatively analyzed, using *ju:mæʧ* scores, a small sample of manually double-annotated Czech data (21 sentences with 255 tokens, annotated by two annotators in parallel). The scores cover all concept instance triples, all relations between nodes, and selected node attributes. To be able to use the same setting for the manually double-annotated data and for the comparison of the manually and automatically created structures, we skip attributes whose values cannot be obtained from the source data (wiki, modal-strength) and not-yet-converted source attributes (mode, polite, quote). The results are shown in Table 2.

UMR node mapping:

corpus	MAN nodes	AUTO nodes	mapped	recall	precision	F$_1$
PDT	375	349	284	76%	81%	78%
PDTSC	442	305	235	53%	77%	63%
PCEDT	307	327	244	79%	75%	77%
total	1124	981	763	68%	78%	72%

Concept and relation comparison (only mapped nodes):*

corpus	MAN triples	AUTO triples	match	recall	precision	F$_1$
PDT	844	819	502	59%	61%	60%
PDTSC	622	633	352	57%	56%	56%
PCEDT	714	588	342	48%	58%	53%
total	2180	2040	1196	55%	59%	57%

Concept and relation comparison:**

corpus	MAN triples	AUTO triples	match	recall	precision	*ju:mæʧ* = F$_1$	*smatch*
PDT	1082	916	502	46%	55%	50%	49%
PDTSC	1318	770	352	27%	46%	34%	37%
PCEDT	916	757	342	37%	45%	41%	51%
total	3316	2443	1196	36%	49%	42%	45%

Table 3: Czech UMRs: quantitative comparison of manual and automatic structures.
(* Unmapped nodes are ignored. ** Unmapped nodes all counted as incorrect.
MAN stands for the manual annotation, AUTO for the automatic conversion.)

The table shows that *ju:mæʧ* was able to successfully map 96% of Czech nodes, with the overall F_1 over 90%. However, it is important to note that these figures were obtained after thorough discussions and reconciliation of problematic cases; as such, they represent an upper bound for what the automatic conversion procedure could achieve. While the inter-annotator agreement is reasonably high in this experiment (though the available data sample is very small), the results still indicate that we cannot expect perfect agreement, given the nature of the UMR framework.

Comparison of manual and automatic UMR structures. A basic quantitative analysis with *ju:mæʧ* scores is provided in Tables 3 and 4. The same setting is preserved (i.e., the scores cover all concept instance triples, all relations between nodes, and the same set of node attributes).

The tables reveal relatively low agreement: only 78% of Czech nodes and 77% of Latin nodes were successfully mapped by *ju:mæʧ*. For these correctly mapped nodes, around 60% of the triples match (57% for Czech and 62% Latin). When all nodes are scored, the overall F_1 drops to 42% for Czech and 51% for Latin. The results are broadly consistent across both languages. For Czech, the most elaborated PDT subcorpus displays consistently better conversion results (*ju:mæʧ* reaching 50%), while the PDTSC subcorpus has a low recall, as discussed in § 4.1.

For comparison, Tables 3 and 4 also provide figures obtained by the *smatch* metric. For both languages, these figures are higher than those of *ju:mæʧ* (increase of 3% for Czech and 10% for Latin). Note that *smatch* uses a different nodes mapping algorithm and that it does not allow for excluding selected attributes.

4 Comparison: Analysis of Differences

Despite rather low results reported above, visual comparison of the graphs for individual sentences often yields a fairly good match. The basic structure typically aligns, and differences are mostly local (concepts, relation types, or local structure).

In this section, we focus on the main sources of disagreement and attempt to determine whether they stem from shortcomings in the conversion process, differing interpretations of the annotation guidelines, or even annotation errors (which can potentially be reconciled). Another possible explanation for the observed results lies in the nature of the UMR framework itself, which—as repeatedly

UMR node mapping:					
MAN nodes	AUTO nodes	mapped	recall	precision	F_1
773	865	629	81%	73%	77%

Concept and relation comparison (only mapped nodes):*					
MAN triples	AUTO triples	match	recall	precision	F_1
1820	1923	1168	64%	61%	62%

Concept and relation comparison:**						
MAN triples	AUTO triples	match	recall	precision	*juːmæʧ* = F_1	*smatch*
2174	2367	1168	54%	49%	51%	58%

Table 4: Latin UMRs: quantitative comparison of manual and automatic structures.
(* Unmapped nodes are ignored. ** Unmapped nodes all counted as incorrect.
MAN stands for the manual annotation, AUTO for the automatic conversion.)

noted in its specification—allows for multiple valid annotations of the same meaning (as the comparison of two manual structures illustrates).

The main differences between automatic and manual UMRs lie in the fact that UMR is more abstract than PDT and, at the same time, allows alternative annotations. In particular, abstract predicates (§ 4.1), event-entity distinction (§ 4.2), and abstract entities (§ 4.3) proved to be challenging.

4.1 Abstract predicates

To foster cross-linguistic comparability of meaning representations, UMR introduces several types of abstract predicates (also called abstract rolesets). Among these, rolesets for nonprototypical predication, so-called implicit rolesets, and predicates for reification need special attention during conversion.

Rolesets for nonprototypical predication. UMR predicates for nonprototypical predication capture possession, location, property and object predication, and identity relationships (e.g., have-91 or belong-91 for possession or have-mod-91 for property predication). In PDT, the corresponding semantic content is represented with the overt verb, typically *být* 'be' or *mít* 'have'.[10] The current version of the conversion keeps the lexical predicates *být* 'be' or *mít* 'have', which, of course, is not in compliance with the UMR specification.

As an exemplification, consider the (shortened) PDT example (1) and its manually and automatically created UMR structures (both simplified). The use of the first abstract predicate have-place-91 does not affect the overall structure at the upper level (the only differences being the node and the relation labels, :ARG0 and :place instead of :ARG1 and :ARG2, respectively). However, the UMR-compliant manual annotation substantially differs from the straightforward PDT annotation when it comes to the representation of the interpersonal relation; it employs the have-rel-role-92 predicate, which captures *sister* as a person (:ARG2) who has a 'sister' relation (:ARG4) to the speaker (:ARG1).

(1) *… je tam sestra…*
'… there is (my) sister there…'

```
MAN:
(b / have-place-91
  :ARG2 (t / place)  'there'
  :ARG1 (p / person
        :ARG2-of (h / have-rel-role-92
                :ARG1 (p2 / person
                      :refer-number singular
                      :refer-person 1st)
                :ARG4 (s / sestra)))) 'sister'

AUTO:
(b / být-011  'be'
    :place (t / tam)  'there'
    :ARG0 (s / sestra))  'sister'
```

Next steps: Typical candidates for nonprototypical predication should be identified: (i) Among the valency frames (framesets) of the verbs *být* 'be' and *mít* 'have', identify those corresponding to UMR predicates for nonprototypical predication, together with adequate argument role mapping. (ii) Determine other candidates for possessive predication (e.g., *vlastnit*, 'own, possess', *patřit (někomu)* 'belong to', possessive pronouns, etc.). (iii) Find relational nouns underlying object predication. However, identification of all candidates for abstract

[10] In these contexts, *být* 'be' or *mít* 'have' are considered predicates, i.e., lexical verbs rather than auxiliaries, in Czech linguistics, with valency frames (PDT analogy to framesets) characterizing each of their senses.

predicates remains a challenging task.

Reifications. Reification, a process of converting a role (= a relation) into a concept, is another important UMR feature. From the conversion perspective, it represents an additional source of disagreement. See, e.g., the manual annotation of example (2), where the :frequency relation is changed to the have-frequency-91 predicate in the manual annotation, while the relation is preserved in the automatic conversion. Formally, the upper structure of the graph is the same, the only changes deal with nodes labeling (the lexical predicate *být-011* 'be' to the reification have-frequency-91) and relations labeling (the role :frequency to :ARG2).

(2) *… teď je to každý rok.*
'… now it's every year.'

```
MAN:
(f / have-frequency-91
    :temporal (t / teď)  'now'
    :ARG1 (e / event)
    :ARG2 (r2 / rate-entity-91
            :ARG3 (t / temporal-quantity
                    :quant 1
                    :unit (r / rok)))) 'year'

AUTO:
(b / být-011  'be'
     :temporal (t / teď)  'now'
     :ARG1 (t2 / ten)  'it (refers to event e)'
     :frequency (r / rok  'year'
                  :mod (k / každý)))  'every'
```

Next steps: Again, while the identification of individual valency frames of *být* 'be', which often underlies such structures, appears to be challenging but doable, automatic recognition of other candidates for reification seems a too ambitious task. As the UMR specification suggests applying reification only if needed, this step can be postponed.[11]

Implicit rolesets. UMR is characterized by a list of implicit rolesets that specify various types of information, the most relevant being the following:

- They can identify meta-language information (e.g., publication-91, hyperlink-91, and street-address-91).
- The second group is formed by predicates that express quantitative observations (e.g., include-91 to represent subsets, as in *some of them*, *23% of voters*; range-91 for *more than 2 months*).
- Yet other implicit rolesets indicate special constructions, as, e.g., comparison (like resemble-91 for *be like John*).
- They can also identify dialog-related structures (e.g., request-confirmation-91 for *Okay?*; say-91 for identifying communication structure (who says what to whom)).

In general, the comparison has revealed that it is very difficult to automatically identify language material in PDT that corresponds to phenomena covered by the implicit rolesets in UMR. Moreover, even if such structures are identified, the use of the relevant implicit roleset typically implies a different structure. Compare, for example, the lower part of (2), with *každý rok* 'every year' specifying frequency; the use of the rate-entity-91 roleset with its :ARG3 role (together with the abstract entity temporal-quantity, see § 4.3 below) makes the structure fairly different.

In particular, abstract predicates indicating meta-language information and those related to dialog structures represent a significant source of differences between the manual annotations and the automatic conversions. As this information is typically not explicitly structured by the language, it is not captured within the deep syntactic annotation (our source data), and thus cannot be straightforwardly converted. This is especially relevant for PDTSC dialogs, as illustrated in (3). The manual UMR structure clearly identifies the speaker and the listener and their role changing through the coreference annotation, in contrast to PDT (and thus to the automatic conversion).

(3) a. *Byla to vaše první motorka?*
'Was this your first motorcycle?'
b. *První.*
'First.'

```
MAN:
(s1s / say-91
  :ARG0 (s1e1 / person :refer-person 1st)
  :ARG2 (s1e2 / person :refer-person 2nd)
  :ARG1 (s1b / have-ord-91
          :quote s1s
          :ARG1 (s1m / motorka  'motorcycle'
                  :ARG1-of (s1b2 / belong-91
                              :ARG2 s1e2)) 'you'
          :ARG2 (s1p / ordinal-entity :value 1)))

(s2s / say-91
  :ARG0 (s2e1 / person :refer-person 1st)
  :ARG2 (s2e2 / person :refer-person 2nd)
  :ARG1 (s2b / have-ord-91
          :quote s2s
          :ARG1 (s2m / motorka  'motorcycle'
                   :ARG1-of (s1b2 / belong-91
                               :ARG2 s2e1)) 'I'
```

[11] A possible way to eliminate this type of disagreement would be to normalize all graphs into reified forms prior to an automatic evaluation.

```
        :ARG2 (s2p / ordinal-entity :value 1)))
:coref ((s1e1 :same-entity s2e2)
        (s1e2 :same-entity s2e1)
        (s1m :same-entity s2m))

AUTO:
(s1b / být-007  'be'
  :ARG1 (s1t / ten)
  :ARG2 (s1m / motorka  'motorcycle'
     :mod (s1e2 / entity :refer-person 2nd) 'you'
     :mod (s1p / první)))  'first'

(s2m / motorka  'motorcycle'
     :mod (s2p / první))  'first'
```

This example illustrates one more open question in the UMR specification: To what extent should UMR annotation reconstruct fragmentary usages and ellipses (highly relevant especially for spoken data and dialogs)? While the complete replay is reconstructed in the manual annotation (= *Motorka to byla moje první.* ‘This was my first motorcycle.’), the PDT annotation, and thus the conversion, is limited to the fragment (= *První motorka.* ‘The first motorcycle.’)

Next steps: Although not explicitly annotated in our source files, meta-language information is also available within the PDT data. The next step, therefore, is to examine the extent to which UMR-relevant data can be extracted and utilized to enhance the conversion process: not only to identify speakers in spoken data but also to recognize elements such as headlines and other pertinent contextual information. Second, more detailed guidelines on proper UMR annotation of fragmentary sentences would improve data consistency.

4.2 Event-related nouns

The UMR specification suggests representing agent nouns as arguments of the respective verbs; thus, for example, *teacher* is a person annotated as :ARG0 participant of the predicate *teach-01*. One might infer that nouns denoting other participants should also be represented with respective eventive concepts (e.g., *food* can be annotated either as a thing being :ARG1 of *eat-01* or just as an instance of the lexical entity *food*). However, it is not clear how far the abstraction should go.

The possibility of multiple correct UMR structures for the same lexical content undermines the potential of any automatic metric considering just one “gold” annotation. It inevitably fails to provide comprehensive insight into the quality of the conversion. Cf. the following text fragment from the beginning of the Czech data (4).

(4) *Vážení čtenáři, …*
‘Dear readers (= subscribers), …’

```
MAN:
(... :vocative (p / person
                :ARG0-of (c / číst-002) 'read'
                :mod (v / vážený))) 'dear'

AUTO:
(... :vocative (c / čtenář  'reader'
                 :mod (v / vážený))) 'dear'
```

Although both structures are correct UMRs, their proper comparison remains a challenge far exceeding the capabilities of a simple automatic metric.

Next steps: Though PDT-scenario does not distinguish which lexemes (words) are related to eventive concepts (verbal predicates) and which are entities, additional language resources can be used to identify at least unquestionable candidates for conversion (as already discussed in Lopatková et al., 2024). In addition, a more detailed specification of the UMR conventions could help reduce the occurrence of such ambiguous cases.

4.3 Abstract entities

Artificial lemmas employed in the PDT-scenario for unexpressed arguments (e.g., #PersPron, #EmpVerb) roughly correspond to UMR basic abstract concepts like person, thing, event. However, since it is not possible to deduce the correct type from PDT and LDT data automatically, the conversion introduces two supertypes: (i) entity, subsuming all UMR non-events (esp. person and thing), and (ii) concept (used esp. in constructions where two or more events, states, or entities are compared). The first supertype is illustrated in ex. (3), where the node s1e2 /person (the possessor) in the manual graph corresponds to the node s1e2 / entity in the converted one.

Further, UMR employs a rich set of abstract entities that identify structured data; for example:

- “entities” (e.g., url-entity, percentage-entity, or ordinal-entity in ex. (3), with the subrole :value),
- “quantities” (e.g., temporal-quantity *každý rok* ‘every year’ in ex. (2)),

In the current version of the conversion procedure, structured data of these types have not yet been processed using abstract entities. Thus, they represent an additional source of disagreement in our comparison. (Semi-)automatic identification of at least most frequent constructions remains one of the important tasks for further improvement.

4.4 Discourse relations

The PDT and UMR schemata represent paratactic structures (such as coordination and discourse relations) in a similar way, by introducing a dedicated node in the graph to represent the whole paratactic construction. As a result, the conversion process is generally straightforward and primarily concerned with technical adjustments. However, when paratactic constructions intertwine with other phenomena (such as relative clauses, represented in UMR as inverted relations) additional complexity arises, making the conversion less trivial. For instance, in example (5) (simplified), the conversion fails to accurately capture the :ARG0-of inverted relations, and the coordinating node and is incorrectly placed one level lower in the graph structure.

(5) *et qui fecere et qui facta aliorum scripsere, multi laudantur.*
'many who have acted, and many who have recorded the actions of others, are praised.'

```
MAN:
(sl / laudo-08   'praise'
   :ARG1 (a / and
      :op1 (p / person
         :quant (m / multus)   'many'
         :ARG0-of (f / facio-02   'act'             : ...))
      :op2 (p2 / person
         :quant m
        :ARG0-of (s / scribo-14   'write, record' : ...))))

AUTO:
(l / laudo-08   'praise'
   :ARG0 (e / entity)
   :ARG1 (m / multus   'many'
      :mod (a / and
         :op1 (f / facio-23   'act'              : ...)
         :op2 (s / scribo-14   'write, record'  : ...))))
```

Next steps: While discourse relations are generally handled correctly, their interaction with more complex constructions will be examined. Conversion will be refined if systematic errors are found.

5 Conclusions

This paper presents a comparison between manually constructed UMRs and those produced by automatic conversion from deep syntactic annotations in existing corpora——specifically, PDT-C for Czech and LDT for Latin. We employed a novel evaluation metric that offers several advantages over existing methods to assess similarity of UMR graphs. The results revealed limitations of the current conversion process, which we further analyzed to suggest areas of possible improvements.

Overall, our evaluation shows that automatic UMR conversion performs comparably for Czech and Latin. However, the analysis also reveals significant challenges inherent to the task, particularly the high level of semantic abstraction required by UMR and the fact that UMR allows for multiple valid representations with varying degrees of granularity. These characteristics complicate both the conversion itself and the evaluation of its accuracy.

Despite the relatively low scores, a simple visual comparison of manual and automatically created graphs often reveals reasonable alignment. This suggests that the automatic procedure—especially after implementing the proposed improvements—could serve as a solid basis for subsequent manual annotation, significantly accelerating and reducing the cost of creating UMR data.

Acknowledgments

The work described herein has been supported by the grants *Language Understanding: from Syntax to Discourse* of the Czech Science Foundation (Project No. 20-16819X) and *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic. It has also been partially supported by the Charles University, GAUK project No. 104924, and SVV project No. 260 821.

The project has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL'98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78. Citeseer.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Julia Bonn, Claire Bonial, Matt Buchholz, Hsiao-Jung Cheng, Alvin Chen, Ching-wen Chen, Andrew Cowell, William Croft, Lukas Denk, Ahmed Elsayed, Eva Fučíková, Federica Gamba, Carlos Gomez, Jan Hajič, Eva Hajičová, Jiří Havelka, Loden Havenmeier, Ath Kilgore, Veronika Kolářová, and 40 others. 2025. Uniform meaning representation 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, and 4 others. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Julia Bonn, Chen Ching-wen, James Andrew Cowell, William Croft, Lukas Denk, Jan Hajič, Kenneth Lai, Martha Palmer, Alexis Palmer, James Pustejovsky, Haibo Sun, Rosa Vallejos Yopán, Jens Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2023a. Uniform meaning representation. LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

Julia Bonn, Skatje Myers, Jens E. L. Van Gysel, Lukas Denk, Meagan Vigus, Jin Zhao, Andrew Cowell, William Croft, Jan Hajič, James H. Martin, Alexis Palmer, Martha Palmer, James Pustejovsky, Zdenka Urešová, Rosa Vallejos, and Nianwen Xue. 2023b. Mapping AMR to UMR: Resources for adapting existing corpora for cross-lingual compatibility. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 74–95, Washington, D.C. Association for Computational Linguistics.

Matthew J. Buchholz, Julia Bonn, Claire Benet Post, Andrew Cowell, and Alexis Palmer. 2024. Bootstrapping UMR annotations for Arapaho from language documentation resources. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2447–2457, Torino, Italia. ELRA and ICCL.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

William Croft and D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge University Press, Cambridge.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

Charles J. Fillmore, Collin Baker, and Hiroaki Sato. 2002. Seeing Arguments through Transparent Structures. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, pages 787–791, Paris. ELRA.

Federica Gamba, Alexis Palmer, and Daniel Zeman. 2025. Boostrapping UMRs from Universal Dependencies for scalable multilingual annotation. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX)*, Vienna, Austria. Association for Computational Linguistics.

Berta Gonzalez Saavedra and Marco Carlo Passarotti. 2014. Challenges in enhancing the Index Thomisticus treebank with semantic and pragmatic annotation. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT-13)*, pages 265–270.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, and 26 others. 2024a. Prague dependency treebank - consolidated 2.0 (PDT-C 2.0). LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 5208–5218, Marseille, France. European Language Resources Association.

Jan Hajič, Eva Fučíková, Markéta Lopatková, and Zdeňka Urešová. 2024b. Mapping Czech Verbal Valency to PropBank Argument Labels. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations (DMR 2024)*, pages 88–100, Torino, Italia. ELRA and ICCL.

R. W. Langacker. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*, volume I. Stanford University Press, Stanford.

Markéta Lopatková, Eva Fučíková, Federica Gamba, Jan Hajič, Hana Hledíková, Marie Mikulová, Michal Novák, Jan Štěpánek, Daniel Zeman, and Šárka Zikánová. 2025. UMR 2.0 - Czech: Release Notes. Technical Report TR-2025-74, ÚFAL MFF UK, Prague, Czechia.

Markéta Lopatková, Eva Fučíková, Federica Gamba, Jan Štěpánek, Daniel Zeman, and Šárka Zikánová. 2024. Towards a conversion of the prague dependency treebank data to the uniform meaning representation. In *Proceedings of the 24th Conference Information Technologies – Applications and Theory (ITAT 2024)*, pages 62–76, Košice, Slovakia. Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia, CEUR-WS.org.

Marco Passarotti. 2014. From syntax to semantics. first steps towards tectogrammatical annotation of Latin. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 100–109, Gothenburg, Sweden. Association for Computational Linguistics.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10:299–320.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.

Haibo Sun and Nianwen Xue. 2024. Anchor and broadcast: An efficient concept alignment approach for evaluation of semantic graphs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1052–1062, Torino, Italia. ELRA and ICCL.

Jens van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, James Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, and Rosa Vallejos. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35(2):343–360.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations (DMR 2023)*, pages 23–33, Nancy, France. Association for Computational Linguistics.

A Node mapping in *juːmæʧ* and *smatch*

Here we show an example sentence from the test data and document the word alignments and the node mapping used by the two metrics.

The full sentence: *Vážení čtenáři, je tomu právě rok, kdy jsme vám oznamovali nepopulární informaci, že se cena našich novin zvyšuje.* "Dear readers, it's been a year since we announced the unpopular news that the price of our newspaper was increasing."

Our excerpt: *Vážení čtenáři, je tomu právě rok, kdy jsme vám oznamovali informaci* "Dear readers, it's been a year since we announced the news"

```
MAN:
(s1p0 / publication-91
  :ARG3 (s1s1 / say-91
    :aspect activity
    :modal-strength full-affirmative
    :ARG0 (s1p1 / person
      :refer-number plural
      :refer-person 1st)
    :ARG2 (s1p2 / person
      :refer-number plural
      :refer-person 2nd
      :ARG0-of (s1c1 / číst-002 'read'
        :aspect habitual
        :modal-strength full-affirmative)
      :mod (s1v1 / vážený 'dear'))
  :ARG1 (s1h1 / have-temporal-91
    :aspect state
    :modal-strength full-affirmative
    :quote s1s1
    :vocative s1p2
    :ARG1 (s1o1 / oznamovat-002 'announce'
      :aspect performance
      :modal-strength full-affirmative
      :ARG0 s1p1
      :ARG1 (s1i1 / informace 'information'
        :refer-number singular)
      :ARG2 s1p2)
    :ARG2 (s1r1 / rok 'year'
      :refer-number singular
      :mod (s1p4 / právě 'just')))))
```

```
AUTO:
(s1b1 / být-011
  :aspect activity
  :vocative (s1c1 / čtenář 'reader'
    :refer-number plural
    :mod (s1v1 / vážený 'dear'))
  :ARG1 (s1t1 / ten
    :refer-number singular
    :temporal (s1p1 / právě 'just'))
  :duration (s1r1 / rok 'year'
    :refer-number singular
    :temporal-of (s1o1 / oznamovat-002 'announce'
      :aspect activity
      :ARG0 (s1p2 / person
        :refer-number plural
        :refer-person 1st)
      :ARG1 (s1i1 / informace 'information'
        :refer-number singular)
      :ARG2 s1c1)))
```

juːmæʧ node mapping between MAN and AUTO (word alignment, if any, is shown in brackets after the concept):

```
s1p0 / publication-91 … UNMAPPED
s1s1 / say-91 … UNMAPPED
s1p1 / person ("našich") … s1p2 / person ("našich")
s1p2 / person ("čtenáři vám")
      … s1c1 / čtenář ("čtenáři vám")
s1c1 / číst-002 … UNMAPPED
s1v1 / vážený ("Vážení") … s1v1 / vážený ("Vážení")
s1h1 / have-temporal-91 ("je") … s1b1 / být-011 ("je")
s1o1 / oznamovat-002 ("tomu jsme oznamovali")
      … s1o1 / oznamovat-002 ("jsme oznamovali")
s1i1 / informace ("informaci")
      … s1i1 / informace ("informaci")
UNMAPPED … s1t1 / ten ("tomu")
s1r1 / rok ("rok kdy") … s1r1 / rok ("rok kdy")
s1p4 / právě ("právě") … s1p1 / právě ("právě")
```

smatch node mapping between MAN and AUTO (showing only differences from *juːmæʧ* mapping):

```
s1s1 / say-91 … s1b1 / být-011 ("je")
s1h1 / have-temporal-91 ("je") … s1t1 / ten ("tomu")
```

In our excerpt, the only nodes left unmapped by *smatch* are s1p0 and s1c1 from the MAN graph, because there are no nodes left available in the AUTO graph. There are two other nodes that are left unmapped by *juːmæʧ* but not by *smatch*: s1s1 in MAN and s1t1 in AUTO. The mapping that *smatch* found for these nodes has no semantic justification (but it will slightly increase F_1 score because both say-91 and být-011 have :aspect activity).